Proceedings Accuracy 2000, Amsterdam, July 2000

Thematic Validation of Land Cover Data of the Eastern United States Using Aerial Photography: Feasibility and Challenges*

Limin Yang[1], Stephen V. Stehman[2], James Wickham[3], Smith Jonathan[3], and Nicholas J. VanDriel[4]

[1]Raytheon, EROS Data Center, Sioux Falls, SD 57198 USA
(Phone. 605-594-6039, Fax. 605-592-6529, Email:lyang@edcmail.cr.usgs.gov)
[2]SUNY ESF, 320 Bray Hall, Syracuse, NY 13210 USA
[3]Environmental Protection Agency, Research Triangle Park, NC 27709 USA
[4]USGS EROS Data Center, Sioux Falls, SD 57198 USA

Keywords: validation, classification, remote sensing

Abstract

Thematic accuracy of a medium spatial resolution (30 meter) land cover data set of the eastern United States was assessed using 1:40,000 scale aerial photos acquired by the United States National Aerial Photography Program (NAPP). The approach implemented for the project proves feasible and cost-effective for validating a large-area land cover product. This paper addresses advantages and limitations in using aerial photos as reference data and presents preliminary results from the data analysis.

1. Introduction

A consortium consisting of several federal agencies was formalized to produce a consistent and seamless land cover base-line product for the conterminous United States (Loveland and Shaw, 1998). Land cover mapping has been conducted for each of ten geographic regions using early 1990s Landsat Thematic Mapper (TM) imagery augmented by a suite of geospatial ancillary data (Vogelmann et al., 1998). This program has provided vital information on national land cover to meet a wide range of user requirements in environmental assessment, monitoring, and modeling at a regional scale.

As land cover mapping of each region is completed, thematic accuracy is evaluated. The accuracy assessment is achieved with a probability sampling design, a response design for reference data evaluation, and an analysis procedure for estimation of accuracy parameters. This validation represents, for the first time, a major effort to rigorously assess the quality of a land cover product for the entire conterminous United States developed from TM imagery.

This paper highlights procedures and preliminary results obtained from an accuracy assessment conducted in four geographic regions over the eastern United States (Fig. 1), including New England (region 1), upper mid-Atlantic (region 2), lower mid-Atlantic (region 3) and southeastern region (region 4). Experiences and challenges in using aerial photos as reference data for large-area land cover assessment are summarized.

* Limin Yang's work was performed under U.S. Geological Survey contract 1434-CR-97-CN-40274. This paper is preliminary and has not been edited or reviewed for conformity with U.S. Geological Survey standards or nomenclature.

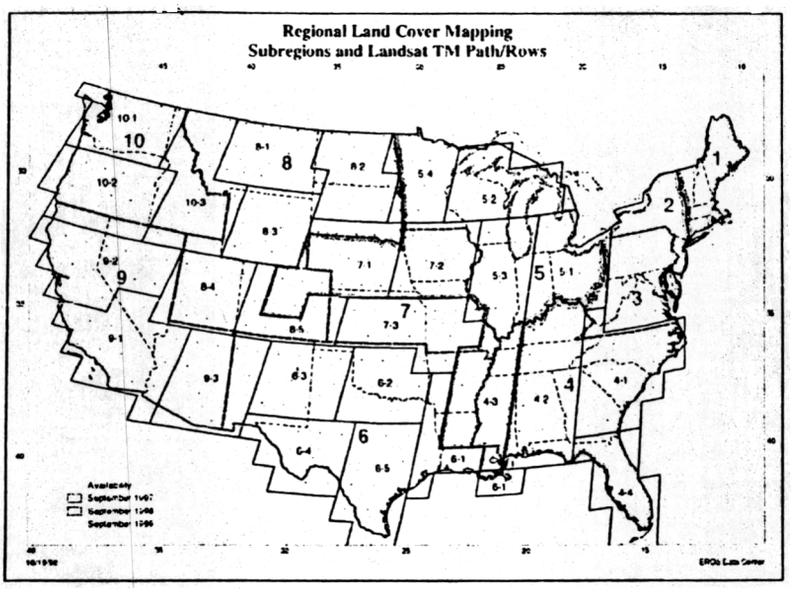

Figure 1. Geographic regions upon which accuracy assessment of the land cover product is conducted.

2. Reference Data Collection

2.1 Reference Data Source

Given national land cover project objectives and resource constraints, NAPP aerial photographs provided the best available reference material for this accuracy assessment. As a national program, NAPP is flown systematically at approximately 5-year intervals over the entire United States, providing an adequate source of reference data from which to design a suitable sampling plan. The NAPP photographs of the early 1990s generally coincide with the date of the Landsat Thematic Mapper (TM) data acquisition used for the land cover classification. Interpretation of 1:40,000-scale aerial photographs is considered a cost-effective way to collect reference data. Limitations to using this approach may include: (1) the often-unavoidable time differences between the TM and NAPP dates, (2) uncertainty in co-registration between TM imagery and photos, and (3) the need for field visits to ascertain photointerpretation.

2.2 Unit of Assessment

Thematic accuracy can be evaluated using a variety of spatial units, including pixel blocks (e.g., 3 x 3 pixels), individual pixels, and polygons (Stehman and Czaplewski, 1998). In this study, pixels were used as the unit of assessment – the same as the basic mapping unit in the final land cover product (unfiltered and unsmoothed). Without accounting for any spatial uncertainty (e.g., mis-registration effects), results of pixel-to-pixel accuracy assessment reflect both misclassification and registration error.

2.3 Response Design Protocol for Reference Data Evaluation

Reference data collection needs to be efficient and consistent across all geographic regions. The response design protocol should take into consideration uncertainties anticipated when interpreting NAPP photos, including: 1) definition of land cover and land use, 2) heterogeneity in land cover characteristics of each sample unit; 3) acquisition dates of Landsat

TM imagery and NAPP photographs; 4) consistency in photo interpretation among interpreters; and 5) sample location.

For each region, sample coordinates were visually transferred from the TM image on the screen to the NAPP photo prints. The sample sites were interpreted directly on the photographs. For each sample unit, the following information was obtained:

- A primary, and possibly one alternate, land cover class (center pixel of a 3 by 3 pixel block)
- Dominant land cover of the 3 by 3 pixel block
- Location of the sample site based on pixel purity, whether the pixel is

1. On the edge of two land cover classes
2. Homogeneous (one land cover class)
3. Heterogeneous (more than two land cover classes)

- Confidence rating of photointerpretation

1. Land cover and land use information is too difficult to interpret
2. Interpretation is perhaps a correct label with some doubt
3. Interpretation is probably a correct label
4. Interpretation is absolutely a correct label

- Notes on any other factors affecting the photointerpretation (e.g., temporal change).

Recording location and confidence rating information allows a better understanding of factors contributing to disagreement between mapped land cover class and reference data. Location informs about the potential impact of mixed pixels, and the confidence rating provides an indicator of the reliability of photointerpretation. Low confidence rating is recorded often because the interpreters cannot determine an appropriate label based on the NAPP photograph. This may be due to an edge condition (e.g., between roads and crop fields), or lack of information on land use (e.g., high intensity residential versus commercial use).

3. Data Analysis and Results

3.1 Consistency in Photo-interpretation:

Ensuring consistency among interpreters is essential to obtaining high quality reference data. Conducting cross-calibration between photo-interpreters as part of an interpreter training program improved, to a certain extent, consistency in land cover labeling and confidence rating. Table 1 and Table 2 display comparisons between two photo-interpreters for lower mid-Atlantic (region 3) and southeastern regions (region 4), respectively. The overall agreement between two photo-interpreters is 79% for region 3 and 84% for region 4. Discrepancies arise primarily in interpreting among three forest classes, and between row crop and hay/pasture. The inconsistencies for forest classes are understood in that mixed forest is often difficult to differentiate from other forest types.

Land cover class	11	21	22	23	31	32	33	41	42	43	81	82	85	91	92	total	% match
11 Open water	137				1			6	1	4		1	1	3		154	89
21 Low int. residential		71		9				1					2			83	86
22 High int. residential																	
23 High int. commercial				62								1	1			64	97
31 Bare rock/sand				1	2		7									10	20
32 Quarry/strip mine						12										12	100
33 Transitional barren		1		1			160	1	2	2	4		4			175	91
41 Deciduous forest		2		1			3	385	21	62	6	2		7		489	79
42 Evergreen forest							2	2	158	21				1		184	86
43 Mixed forest		1		1			14	20	40	202		5	1	16		297	68
81 Hay/pasture				2			1				101	10	11			125	81

82 Row crops				2				2		1	18	175	1			199	88
85 Other grass (in urban)		7		5		13	17	4		20	6		76			148	51
91 Woody wetlands								6		1			2	157		166	95
92 Emergent wetlands								1		1			5	6	56	69	81
total	137	82		84	3	25	204	428	222	314	135	191	104	190	56	2175	
% match	100	87		74	67	48	78	90	71	64	75	92	73	83	100		79

Table 1. Comparison of consistency between photo-interpreters, region 3 (Data from Roth, et al., 1999).

Land cover class	11	21	22	23	31	32	33	41	42	43	81	82	85	91	92	total	% match
11 Open water	18	1												1		20	90
21 Low int. residential		21											1			22	95
22 High int. residential			3	1												4	75
23 High int. commercial				9												9	100
31 Bare rock/sand					4							2				6	67
32 Quarry/strip mine	1					6										7	88
33 Transitional barren							16	1					1			18	89
41 Deciduous forest							2	14		1	1	1				19	74
42 Evergreen forest							2		7							9	80
43 Mixed forest							3	2	1	26				4		36	72
81 Hay/pasture											10	1				11	91
82 Row crops											3	10		1		14	71
85 Other grass (in urban)								1					16		2	19	84
91 Woody wetlands										2				13	1	16	81
92 Emergent wetlands															15	15	100
total	19	22	3	10	4	6	23	18	8	29	14	14	18	19	18	225	
% match	95	95	100	90	100	100	70	78	88	90	71	71	89	68	83		84

Table 2. Comparison of consistency between photo-interpreters, region 4. (Data from Khorram, et al., 1999)

On the other hand, cropland and hay/pasture are difficult to distinguish with just a single date aerial photo due to crop rotation practices. A problem also exists for the urban grass class (e.g. parks, golf course) in region 3. This class was confused with woodland and other urban residential classes. Final labels for sample points in which disagreement between two interpreters occurred were decided through a group discussion based on characteristics of the sample sites and its neighboring pixels. In some cases the difference can be resolved through a consistency check among different interpreters. In other cases no best label could be interpreted from the photo so a compromise had to be made and a low confidence rating assigned to the sample unit.

3.2 Results from Different Definition of Agreement

Accuracy results can be reported using several definitions of agreement between the map and reference labels. The rationale of using different definitions is to bracket the map accuracy between optimistic and conservative estimates. A direct comparison at each pixel of the photo-interpreted land cover label with the corresponding map label is the most sensitive protocol for defining agreement. The results of this comparison are most affected by confounding non-mapping errors such as image geo-registration uncertainty. The second definition of agreement allows a match between the photo-interpreted sample pixel and a dominant class mapped within a 3 by 3 pixel block centered on the sample pixel. This comparison assumes that, for many applications, a certain level of generalization from the full resolution (30 meters) land cover data is adequate. The third definition of agreement allows a match if the reference label agrees with the map label for any one of the 9 pixels forming the 3x3 block centered on the sample pixel. This definition allows a map shift of one pixel in any direction to account for location uncertainty resulting from either the image registration

and/or from locating pixels on aerial photo. This is the least sensitive definition of agreement and provides an upper bound on accuracy.

Using different definitions of agreement, a significant change was noted in the proportion of agreement between mapped and reference land cover classes. Taking the southeast region as an example, according to a report by Khorram et al., (1999), the overall agreement changed from 55.9% (pixel-to-pixel comparison, including the possible alternate label) to 66.8% (dominant class within a 3 by 3 pixel block), to 79% (any pixel within a 3 by 3 pixel window). Results of individual classes also changed. The agreement of the urban low intensity residential class, for instance, increased from 44.5% (pixel-to-pixel) to 80% (dominant class) to 87% (any pixel). A similar pattern holds for mixed forest class; the proportion of agreement increased from 43.0% (pixel-to-pixel) to 64.9% (dominant class) to 82% (any pixel). The other grass class (parks and golf courses in urban area) also increased from 46.6% (pixel-to-pixel) to almost 70% (any pixel). The increase in agreement indicates that these classes are likely correctly mapped in terms of spatial context even though not all individual pixels are mapped correctly.

3.3 Results using Different Confidence Sites

As was discussed in section 2.3, a confidence rating index was assigned to each sample unit by photo-interpreters. Comparing accuracy estimates obtained from using only high confidence samples versus using all sample sites allows insights on the effect of several factors on accuracy estimates, including land use/cover definition, mixed pixels, time changes between the date of imagery and the date of the NAPP photo, and disagreement among photointerpreter.

Table 3 shows, using a subset data from the region 2 (upper mid-Atlantic), a much higher percentage of agreement for sample pixels in homogeneous areas or with a high confidence rating. Table 4, using the same sample data set, indicates a positive correlation between confidence and location rating. In general 86 percent of samples with high confidence rating are the ones in homogeneous areas, whereas 64 percent of samples with low confidence rating are in areas of mixed land cover types. The fact that accuracy increases for the high confidence and/or homogeneous sample sites suggests that limiting accuracy sampling to clearly interpretable (homogeneous) pixels would have provided a much more optimistic view of accuracy.

Confidence Rating	**Number of pixels**	**Number in agreement Map versus photos**	**% agreement**
Low (2)	47	14	29.7
Moderate (3)	142	80	56.3
High (4)	271	181	66.7
Location Rating	**Number of pixels**	**Number in agreement Map versus photos**	**% agreement**
Edges (1)	151	81	53.6
Homogeneous (2)	282	182	64.5
Heterogeneous (3)	27	12	44.0

Table 3. Classification agreement based on different confidence and location rating. Sample points (460) are obtained from the New York/New Jersey region. Percentages are based on row totals.

	Homogeneous (2)		Heterogeneous (3)		Edges (1)	
Confidence Rating	Number	% of total	Number	% of total	Number	% of total
Low (2)	6	27	14	64	2	9
Moderate (3)	45	29	13	8	99	63
High (4)	242	86	0	0	38	14

Table 4. Relation between confidence rating and location rating, Sample points (460) are obtained from the New York/New Jersey region. Percentages are based on row totals.

3.4 Most Frequently Confused Classes

Table 5 lists the most frequently confused land cover categories (mapped versus photo-interpreted) for four regions. This table helps to understand to what extent the confused classes are among similar land cover types. Most often confusion occur between related classes, i.e., among the three urban land use classes, among the three forest classes, and between row crop and hay/pasture. However, a few land cover types are confused with many other classes. Transitional barren, defined as areas dynamically changing from one land cover to another because of land use activities, is an example of such a class. Another problem noted is the confusion between two barren classes and forest and grassland classes in the mid-Atlantic and southeast region.

3.5 Factors Affecting Land Cover Mapping

For the eastern United States, water, urban, and forest classes are generally mapped well, whereas forested wetland, transitional barren, hay/pasture and crops have more confusion. Factors that have contributed to the confusion can be categorized into those associated with: 1) mapping error, 2) timing of reference data acquisition (hay/pasture, row crop, wetland, transitional), 3) definition related to land use (high intensity residential and urban built-up, and the two barren classes), and 4) geo-registration error.

Map class name	Region 1	Region 2	Region 3	Region 4
Open water			Emergent wetland	Emergent wetland
Low int. residential	Evergreen forest	High int. residential	Other grass	Mixed forest
High int. residential	High int. commercial	High int. commercial	High int. commercial	Low int. residential
High int. commercial	Low int. residential	Low int. residential	Low int. residential	Low int. residential
Bare rock/sand	Shrublands	Emergent wetland	Mixed forest	Other grass
Quarry/strip mine	High int. commercial	High int. commercial	Deciduous forest	Bare rock/sand
Transitional barren	Shrublands	Woody wetlands	Evergreen forest	Mixed forest
Deciduous forest	Mixed forest	Mixed forest	Mixed forest	Mixed forest
Evergreen forest	Mixed forest	Deciduous forest	Mixed forest	Mixed forest
Mixed forest	Evergreen forest	Evergreen forest	Evergreen forest	Transitional barren
Hay/pasture	Fallow	Row crops	Row crops	Row crops
Row crops	Hay/pasture	Hay/pasture	Deciduous forest	Hay/pasture
Other grass	High int. residential	Low int. residential	High int. commercial	Hay/pasture
Woody wetlands	Evergreen forest	Evergreen forest	Deciduous forest	Mixed forest
Emergent wetland	Evergreen forest	Woody wetlands	Open water	Woody wetlands

Table 5. The most frequent confusion between mapped and photointerpreted land cover classes by region.

An example of mapping error is the limitation in using leaf-off season (spring or fall) Landsat TM data for discriminating between hay/pasture and row crops. An assumption is that there is a temporal window during which hay and pasture areas green up before most other annual or perennial vegetation. However, if leaf-off data are not temporally ideal (e.g., the greenness level of hay/pasture areas is low), then errors involving confusion between hay/pasture and other agricultural lands will result.

Acquisition dates of the NAPP photographs used as reference data range from the late 1980s to 1997, whereas the satellite data vary mostly from 1991 to 1993. The changes that have taken place across the landscape over this time can complicate interpretation and comparison. One of the major problem classes is transitional barren, a class that is designed for conditions such as temporary clearing and regeneration of forest cover.

Low accuracy for classes that relate to land use in nature is understandable. Despite the extensive use of ancillary data, such as the population census, it is very difficult to unambiguously separate high intensity residential from other urban use, either during the modeling of TM data or simply when viewing it on a NAPP photograph. The same is true for the land use related differences between the quarry/strip mine class and the sandy/gravel class.

4. Conclusions

A land cover data set of the eastern United States was evaluated using NAPP aerial photos. The procedure developed for this exercise generally worked well using an appropriate sampling design (Stehman, et al., 2000, this issue) and a response design for reference data evaluation. Because of several constraints related to using aerial photos as reference data for this accuracy assessment, we examined the quality of land cover data through a comparison over a range of definitions for agreement. This approach is likely to present a better perspective on the quality of the land cover data set being assessed.

Several challenges have been encountered in this project. First, time differences between the TM and NAPP photo dates present difficulty in direct comparison between mapped land cover class and reference data. This issue has not yet been examined and needs further investigation. Another issue concerns separating location error from mapping error, which is not an easy task given the current procedure used in photo-interpretation. Spatial uncertainty of a given sample pixel can arise from two sources: those associated with geometric accuracy of satellite imagery (+/- 30 meters), and those related to locating sample units from satellite data on non-georeferenced NAPP photos. Either source of uncertainty could adversely impact accuracy estimates. Yet another important issue is the inconsistency among photo-interpreters. The challenge is in quantifying this effect and accounting for it in deriving accuracy estimates.

Reference

Khorram, S., Di, X., Knight, J., Yuan, H., Cakir, H., and Mao. Z., 1999, Accuracy assessment of the EPA region VI dataset of the MRLC land cover mapping program. Final report. Prepared by Center for Earth Observation, North Carolina State University, Raleigh, North Carolina, USA.

Loveland, T.R., and Shaw, D.M., 1996, Multi-resolution land characterization--building collaborative partnerships, in Scott, J.M., Tear, T.H., and Davis, F.W., eds., GAP analysis--a landscape approach to biodiversity planning: Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing, p.79-85.

Roth, N., Strebel, D., Kou, J., and Krebs, T., 1999, An assessment of land cover data used in the Mid-Atlantic landscape atlas. Final report. Prepared by Versar, Inc., Columbia, MD for the U.S. Environmental Protection Agency.

Stehman, S.V., and Czaplewski, R.L., 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, 64:331-344.

Stehman, S.V., Wickham, J., Yang, L., and Smith, J., 2000, Assessing accuracy of large-area land cover maps: Experiences from the Multi-Resolution Land-Cover Characteristic Project, (this issue)

Vogelmann, J.E., Sohl, Terry, and Howard, S.M., 1998, Regional characterization of land cover using multiple sources of data: *Photogrammetric Engineering and Remote Sensing.* 64: 45-57.

Zhu, Z., Yang, L., Stehman, S.V., and Czaplewski, R.L., 2000, Accuracy assessment of the USGS regional land cover characterization, New York and New Jersey. *Photogrammetric Engineering and Remote Sensing*, (in press).